Boffin Access

Open Access Journal | Volume: 1.2

Advances in Computer Sciences

ISSN 2517-5718

Augmenting Activity Recognition with Commonsense Knowledge and Satellite Imagery

Nicola Bicocchi[1*]
Damiano Fontana[2]
Franco Zambonelli[2]

[1]Department of Information Engineering, University of Modena and Reggio Emilia Modena, Italy
[2]Department of Sciences and Methods of Engineering, University of Modena and Reggio Emilia, Italy

Abstract

Activity recognition gained relevance because of its applications in a variety of fields. Despite relevant improvements, classifiers are still inaccurate in several real-world circumstances or require excessively time-consuming training routines. In this paper we show how satellite imagery and common sense knowledge can be used for improving users' activity recognition performed on a mobile device. More specifically, we made use a personal device providing a list of candidate user activities instead of only the most probable one. Then, from the GPS location of the user, we (i) extract a list of neighboring commercial activities using a reverse geo-coding service and (ii) classify the satellite imagery of the area with state-of-the-art techniques. The proposed approach uses the ConceptNet network for ranking the list of candidate activities using both additional information. Results show an improvement in activity recognition accuracy.

Keywords

Pervasive Computing; Activity Recognition; Mobility; Commonsense Knowledge

Introduction

Pervasive systems, constantly analyzing different facets of our world, frequently cooperate to provide services with a coherent representation of the environment. However, despite the many facets of our life are strictly tied from the practical viewpoint (e.g., if a user is running he is likely to be in suitable location such as a park or a gym), it is difficult to exploit their correlation using traditional learning techniques (e.g., bagging, boosting) [1]. On the other side, treating each facet as an independent variable might lead to unrealistic results. For instance, locations and activities are strictly correlated.

In this paper we tackle the problem of enabling situation-recognition capabilities by fusing different sensor contributions. Specifically, we propose to extract well-known correlations among different facets of everyday life from a commonsense knowledge base. The approach is general and can be applied to a number of cases involving commonsense for the sake of: (i) ranking classification labels produced by different classifiers on a commonsense basis (e.g., the action classifier detects that the user is running with an high confidence and the place classifier outputs two possible labels: "park" and "swimming pool". In this case, using commonsense, it is possible to infer that the user is more likely to be in a park that in a swimming pool); (ii) predicting missing labels (e.g., if a user is running but the location data is missing, it is possible to propose "park" as a likely location). More in details, the paper contains three main contributions: (i) it describes a greedy search algorithm to measure the semantic proximity of two concepts within the ConceptNet network [2]; (ii) it proposes a novel technique to extract contextual localization data from satellite imagery; and (iii) it shows how to improve activity recognition accuracy by making use of two different localization sensors and common sense knowledge.

Accordingly, the rest of the paper is organized as follows: Section II formally defines the problem of commonsense sensor fusion and describes the proposed algorithm. Section III describes the experimental testbed we implemented to validate our proposal. Section IV details experimental results under different configurations. Section V discusses related work. Finally, Section VI concludes the paper.

Sensor Fusion with Commonsense Knowledge

The proposed approach is based on the assumption that commonsense knowledge can be used to measure the semantic proximity among concepts. The more two concepts are proximate, the more it is likely they have been recognized within the same context [3]. In this section we formally introduce the approach.

A. Problem Definition

Let us consider a set of n classifiers C_1::C_n, each one delegated to recognize a specific facet of the environment. Each classifier C_x is able to deal with uncertainties by producing (at every time step t) m labels $l_1(C_x; t)$; :::; $l_m(C_x; t)$ for each data sample. Given that, the

Article Information

DOI:	10.31021/acs.20181110
Article Type:	Research Article
Journal Type:	Open Access
Volume:	1 **Issue:** 2
Manuscript ID:	ACS-1-110
Publisher:	Boffin Access Limited
Received Date:	February 21, 2018
Accepted Date:	April 20, 2018
Published Date:	May 08, 2018

***Corresponding author:**

Nicola Bicocchi
Department of Information Engineering
University of Modena and Reggio Emilia
Modena, Italy
Tel. No: +393478004903
E-mail: nicola.bicocchi@unimore.it

Citation: Bicocchi N, Fontana D, Zambonelli F. Augmenting Activity Recognition with Commonsense Knowledge and Satellite Imagery. Adv Comput Sci. 2018;1(2):110

overall perception of the environment can be represented as a tuple $((l_1(C_1; t); :::; l_m(C_1; t)); :::; (l_1(C_n; t); :::; l_m(C_n; t)))$.

In this paper, we tackle the problem of ranking all the possible tuples provided by n classifiers on a commonsense basis.

The general problem of commonsense tuple ranking can be expressed, without loss of generality, in this way: given 2 tuples both composed by commonsense concepts, $(l_1(C_1; t); l_1(C_2; t))$ and $(l_2(C_1; t); l_2(C_2; t))$, is it possible to establish which tuple contains the most proximate concepts on a commonsense basis?

Measuring commonsense proximity requires two key conditions to be met. In particular: (i) a knowledge base containing both a vocabulary covering a wide scope of topics and semantic relations hard to be discovered in an automatic way; and (ii) an algorithm for computing semantic proximity.

The first condition is best addressed by ConceptNet. It is a semantic network designed for commonsense contextual reasoning. It was automatically built from a collection of 700,000 sentences, a corpus being a result of collaboration of some 14,000 people. It provides commonsense contextual associations not offered by any other knowledge base. ConceptNet is organized as a massive directed and labeled graph. It is made of about 300,000 nodes and 1.6 million edges, corresponding to words or phrases, and relations between them, respectively. Most nodes represent common actions or chores given as phrases (e.g., “drive a car” or “buy food”). Its structure is uneven, with a group of highly connected nodes, and “person” being the most connected, having in- degree of about 30,000 and out- degree of over 50,000. There are over 86,000 leaf nodes and approximately 25,000 root nodes. The average degree of the network is approximately 4.7.

To meet the second requirement, we started from a preliminary round of experiments with ConceptNet that led us to the following principles:

1) Proximity increases with the number of unique paths. However, this is not a reliable indicator given that even completely unrelated concepts might be connected through long paths or highly connected nodes.
2) Proximity decreases with the length of the shortest path; nodes connected directly or through some niche edges are in a short distance, hence they are proximate;
3) Connections going through highly connected nodes increase ambiguity, therefore proximity should be inversely proportional to the degrees of visited nodes;
4) ConceptNet has been created from natural-language assertions. Thus, errors are frequent and algorithms have to be noise-tolerant;

Majewski et al. recently proposed an interesting algorithm for commonsense text categorization inspired by similar observations [9]. Despite having been conceived for a different problem, it can be applied to localization as well. The algorithm is based on the assumption that proximity among concepts is proportional to the amount of some substance s that reaches the destination node v as a result of injection to node u. The procedure has been built around two key biological paradigms such as diffusion and evaporation and works as follow:

1) a given amount of substance s is injected to a node u;
2) at every node, a fraction of the substance evaporates and leaves the node;
3) at every node, the substance diffuses into smaller flows proportional to the out degree of the node;
4) Nodes never overflow. If multiple paths visit the same node, the previous amount of substance s can be incremented;
5) Target nodes are ranked according to the amount of substance s received.

Figure 1 exemplifies the algorithm in action. A certain amount (i.e., 256 units) of substance s is injected into a node (i.e., Run). Then, the substance diffuses over the graph and halves by evaporation at each node it visits. The amounts of s that reach nodes Park and Road are 60 and 16 respectively. Park is considered more proximate than Road to Run.

It is worth noticing that this approach can easily handle the fact that different classifiers might produce the same set of labels (i.e., classifiers observing the same facets of reality). In fact, if a label compares multiple times it is sufficient to multiply the amount of substance injected into the corresponding nodes. Furthermore, this approach permits to assign different weights to different classifiers in a straightforward way.

Finally, it is interesting to note how this algorithm matches with the principles we deduced from our preliminary studies on ConceptNet. In fact: (i) the evaporation process assures that short paths imply high proximity; while (ii) the diffusion process takes into account the total amount of connections among two concepts while diminishing the relevance of highly-connected paths.

In the following, we apply the described technique to fuse information contributions coming from sensors analyzing different facets of the same situation.

Improving Activity Recognition

To assess the relevance of our ideas, we used a specific instance of the general problem. We prototyped a system able to improve activity recognition accuracy by making use of two different localization sensors. Activities are classified from accelerometer data while locations from GPS traces. All three modules have been configured to eventually produce multiple labels to deal with uncertainties. In these cases, common sense reasoning is applied.

A. Activity Recognition

To classify user’s activities we implemented a sensor based on [4]. It collects data from 3-axis accelerometers, sampling at 10Hz, positioned in 3 body locations (i.e., wrist, hip, ankle) and classifies activities (i.e., dance, use stairs, drive, walk, run, stand still, drink) using instance-based algorithms. Furthermore, considering that human activities have a minimum duration, it aggregates classification results over a sliding window and performs majority voting on that window. Each window is associated with the most frequent label. For the sake of the experimentation, we modified it to deal with uncertainties. Instead of producing a single label for each sensor sampling, we implemented a mechanism to produce multiple labels associated with a degree of confidence. Specifically, for each sample to be classified, k nearest neighbours (associated to q classes, k = 64, q <= k) are identified. The sample is then associated to all the classes (at most 3) associated to at least k=2q training samples. Table 1 reports a realistic confusion matrix for this sensor.

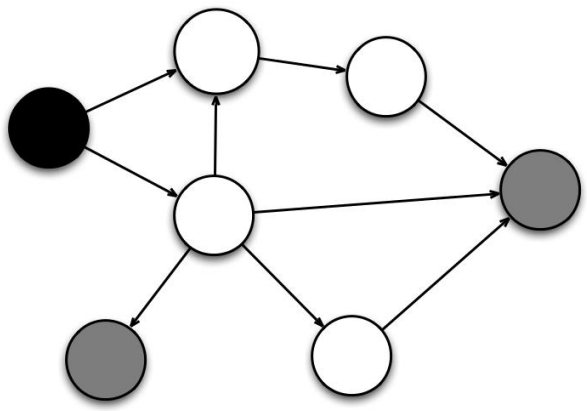

Figure 1: Concept proximity algorithm in action. 256 units of substances are injected into node Run. Then, the substance diffuses over the graph and halves by evaporation (= 0:5) at each node it visits. The amounts of s that reach nodes Park and Road are 60 and 16 respectively. Park is considered more proximate than Road to Run.

B. Location Recognition via Satellite Imagery

The location sensor based on satellite images is based on GPS data and classifies user location by making use of satellite imagery. Specifically, given the GPS coordinates, it uses Google Maps API to retrieve the corresponding image tile. Then, it classifies the tile against a set of 5 categories (i.e., green, harbour, parking, rail, residential). (Figure 2)

To implement this sensor we used an approach based on the bag-of-features image classification technique [5]. In computer vision, the bag-of-words model (BoW model) can be applied to image classification, by treating image features as words. In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features.

A dataset comprising 200 tiles, evenly distributed among the 5 categories, has been collected from Google Maps and manually annotated. SURF features, chosen because of their robustness to scaling and rotation, have been extracted [5].

SURF features have been organized in bag-of-words representing each of the categories and used to train separate one-class SVM classifiers. During the testing stage, instead, the tile covering the user location is downloaded and tested against each classifier. The tile is assigned to the category associated with the classifier with the highest likelihood. Table 2 shows the confusion matrix. All the classes are recognized with an accuracy comprised between 78% and 95%.

C. Location Recognition via Reverse Geocoding

The location sensor based on reverse geocoding samples GPS coordinates and classifies user's location by querying the reverse geocoding Google Maps API [6]. Specifically, this API takes as input the GPS coordinates and a search radius and returns a list of points of interest associated to a label coming from a predefined set (i.e., road, square, park, shop, cinema, mall, restaurant, gym). Unfortunately several practical drawbacks affect this process. Google Maps database, for example, is not perfect. Although we do not have accurate statistics, we noticed that a portion of locations is still missing. Furthermore, locations' coordinates are not always precise. Finally, Google Maps does not provide in-formation about locations' geometry. Due to this, especially for large-sized instances (e.g., parks, squares) locations can be misclassified. For example, a user running close to the border of a park is likely to be associated to the shops she is facing instead of to the park itself.

To mitigate these problems and avoid false negatives, the system has been setup to use a search radius of 250m. Clearly, the number of reverse geo-coded locations is proportional to the search radius. Because of this, especially in densely populated areas, the system might produce numerous false positives. To reduce them, while keeping an acceptable level of false negatives, we implemented 3 filters acting on the GPS signal. Specifically: the first acts on the assumption that each class is more likely to be visited during defined portions of the week. The second, acts on the assumption that each class of locations is fairly characterized by the duration of the visit. This duration is usually related with a GPS signal interruption. Finally, the third one, filters out each label not compatible with measured speed.

Experimental Evaluation

To assess the feasibility of our idea, we used the system described in Section III to collect a dataset comprising a full day of a single user.

The activity recognition module, has been trained to classify 8 activities (i.e., climb, use stairs, drive, walk, read, run, use computer, stand still, drink). For each class, 300 training samples have been selected. The location module implementing reverse geocoding, instead, sampled GPS coordinates each 30 seconds. GPS coordinates has been labeled with 5 different categories (i.e., street, university, bar, park and library).

We first discuss the performance of recognition modules, considered independently. Figure 3(a)(b)(c) summarizes the results. The reverse geocoding localization sensor is the less precise among the three. It correctly recognizes only 20% of locations because of multiple commercial activities are usually located within its search radius. On the other hand, the satellite-based sensor correctly classifies around 80% of the samples. Finally, the activity recognition module is around 65% of correctly classified samples.

When both location and activity labels are combined using ConceptNet, 4 cases can occur: (i) both are available, (ii) only activity is available, (iii) only location is available, (iv) no data available. The first case allows applying common-sense sensor fusion. In both the second and the third case, instead, commonsense can be used to identify a possible place or activity to complete the (activity, place) tuple.

Figure 3(d)(e) show the results obtained by combing activity labels with both the location labels. A significant improvement has been achieved. It is worth noticing that the Undefined (i.e., multiple labels available) category is lowered to zero meaning that ConceptNet is always capable of providing a ranking of action-place couples. Furthermore, the No Classification data category is lowered to zero, in fact one of the advantage of the use of ConceptNet is to provide missing data. Please note that in our experiment we never

experienced the concurrent lack of both sensorial data that should have called for different strategies similar to activity and location prediction, such as bayesian networks [7]. It is worth noticing that the fusion process with the satellite-based sensor produced better results because of its initial performance was better than the reverse geocoding one.

	Dance	Stairs	Drive	Walk	Run	Stand	Drink
Dance	0.89	0	0	0	0.11	0	0
Stairs	0	0.76	0	0.1	0.14	0	0
Drive	0.17	0	0.83	0	0	0	0
Walk	0.1	0	0	0.72	0.18	0	0
Run	0.12	0	0	0	0.88	0	0
Stand	0	0	0	0	0	0.91	0.09
Drink	0	0.1	0	0	0	0.17	0.73

Table 1: Confusion Matrix of the Activity Recognition Sensor

	Park	Harbour	Parking	Rail	Residential
Park	0.89	0	0.01	0	0.1
Harbour	0.07	0.86	0	0	0.07
Parking	0	0	0.78	0	0.22
Rail	0	0	0.05	0.95	0
Residential	0.06	0.03	0.05	0.05	0.81

Table 2: Confusion Matrix of the Localization Sensor Based on Satellite Imagery

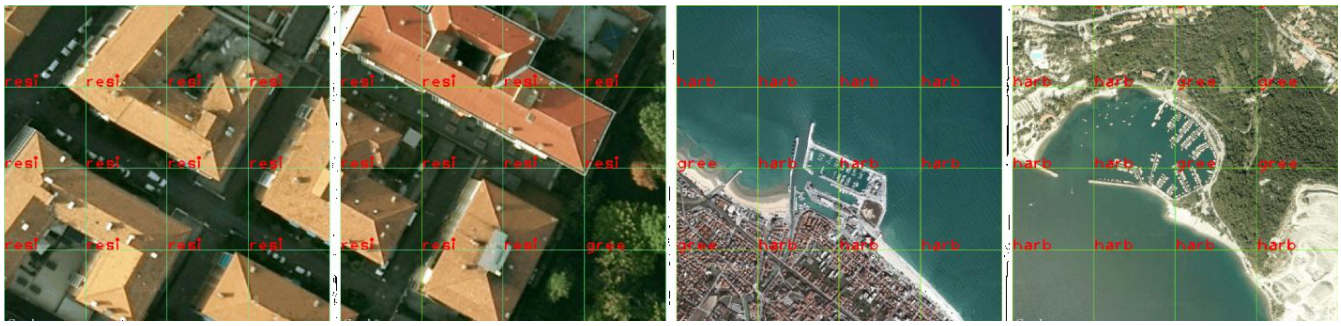

Figure 2: Four snapshot taken from the location sensor using satellite imagery. Residential and harbour areas are correctly classified. The red strings superimposed on map tiles are actual classification labels produced by the sensor.

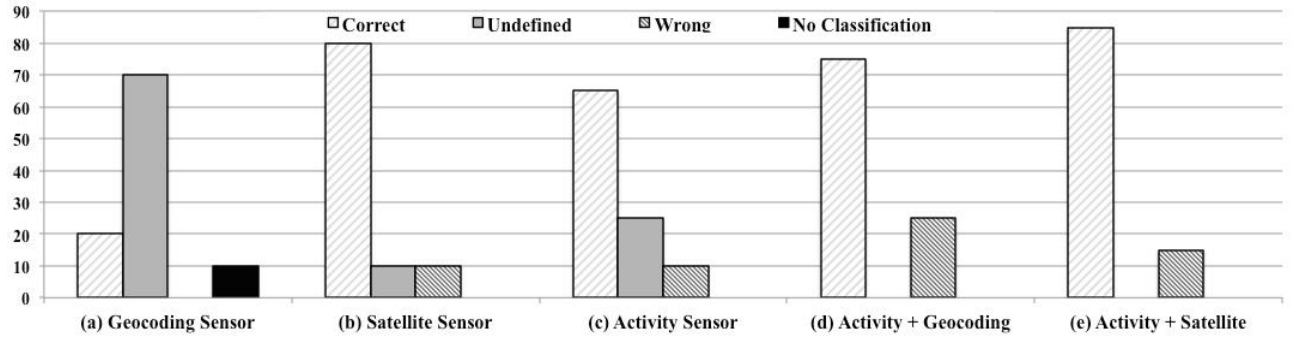

Figure 3: The reverse geocoding sensor correctly recognizes only 20% of locations (a). On the other hand, the satellite-based sensor correctly classifies around 80% (b). Finally, the activity recognition module is around 65% of correctly classified samples (c). Figures (d) and (e) show the results obtained by combing activities with locations coming from both the localization sensors.

Related Work

Many works focus on data fusion at different levels, either for acquiring and making accessible diverse contextual aspects or for reasoning about them. The traditional approach makes use of probabilistic models. [8] Proact combines data coming from RFIDs and an accelerometer mounted on the RFID glove in order to identify activities. RFID tags are used to restrict the number of possible actions by considering the manipulated object. In a system for multi-modal sensor fusion specifically designed for smartphone is proposed [9]. The system exploits data coming from the microphone and inertial sensors on the mobile for inferring high level activities with light-weight bayesian learning algorithms.

Few works make use of commonsense for situation recognition. An interesting approach is presented in [10]. It uses RFID to trace a set of everyday objects and infers user activities by making use of Google searches. Alternatively, applies commonsense to localization [11]. It uses Cyc to improve automatic place identification on the basis of user historical data. However, both these approaches limit the use of commonsense to improve a single contextual aspect. Alternatively, in this paper we use commonsense to integrate multiple aspects.

To best of our knowledge there are few works using common sense to integrate different context sources. Pentland et. al. [12] presented a user-centric situation recognition system able to overhear users' conversations and use ConceptNet as reasoning system. Bicocchi et al. [13], instead, presented a workflow to classify a situation using a stream of images collected from ego-vision devices. Images are independently classified using k-nn search are combined together using commonsense. However, they both do not make use of commonsense knowledge to fuse different contributions.

Conclusion

In this paper we presented preliminary results we obtained with a novel approach that combines an activity classifier and location classifier using satellite imagery with the ConceptNet knowledge base. Different classifiers are fused together on a commonsense basis for both: (i) improve classification accuracy and (ii) dealing with missing labels. The approach has been discussed through a realistic case study focused on the recognition of both locations visited and activities performed by a user. Results have been encouraging and apparently indicates that our approach can be applied to different scenarios.

Acknowledgment

Work supported by the ASCENS project (EU FP7-FET, Contract No. 257414).

References

1. S Bandini, A Mosca, M Palmonari. Common-sense spatial reasoning for information correlation in pervasive computing. Applied Artificial Intelligence. 2007;21(4-5):405–425
2. H Liu, P Singh. Conceptnet, a practical commonsense reasoning tool-kit. BT Technology Journal. 2004;22(4):211–226
3. P Majewski, J Szymanski. Text categorization with semantic commonsense knowledge: First results. In M Ishikawa, K Doya, H Miyamoto, T Ya-makawa. Neural Information Processing. 2008;pp:769–778
4. N Bicocchi, M Mamei, F Zambonelli. Detecting activities from body-worn accelerometers via instance-based algorithms. Pervasive and Mobile Computing. 2010;6(4):482–495
5. H Bay, A Ess, T Tuytelaars, L Van Gool. Speeded-up robust features (surf). Computer Vision and Image Understanding. 2008;110(3):346–359
6. L Ferrari, M Mamei. Discovering daily routines from Google latitude with topic models. In Proceedings of 11th IEEE International Conference on Pervasive Computing and Communications Workshops. 2011;pp:397–402

7. N Bicocchi, G Castelli, M Mamei, A Rosi, Zambonelli. Supporting location-aware services for mobile users with the whereabouts diary. In: Proceedings of the 1st International Conference on Mobile Wireless Middleware, Operating Systems, and Applications. 2008;6:1–6
8. M Stikic, T Huynh, K Van Laerhoven, B Schiele. ADL recognition based on the combination of RFID and accelerometer sensing. In: 2nd International Conference on Pervasive Computing Technologies for Health-care. 2008;pp:258–263
9. RK Ganti, S Srinivasan, A Gacic. Multisensor fusion in smartphones for lifestyle monitoring. In Proceedings of the 2010 International Conference on Body Sensor Networks. IEEE Computer Society. 2010;pp:36-43
10. M Philipose, K Fishkin, M Perkowitz, D Patterson, D Fox, et al. Inferring activities from interactions with objects. IEEE Pervasive Computing. 2004;3(4):50–57
11. M Mamei. Applying commonsense reasoning to place identification. IJHCR. 2010;1(2):36–53
12. A Pentland, T Choudhury, N Eagle, P Singh. Human dynamics: computation for organizations. Pattern Recognition Letters. 2005;26:503–511
13. N Bicocchi, M Lasagni, F Zambonelli. Bridging vision and commonsense for multimodal situation recognition in pervasive systems. In PerCom. 2012;pp:48–56